Match Results for TRI and NETS

Henry Chen

July 17, 2023

Contents

1 Introduction

This documentation is a reference of the Matching between National Establishment Time Series (NETS) and Toxics Release Inventory (TRI) Datasets. You can download the raw data of TRI from the U.S. Environmental Protection Agency Website. The link to the TRI basic data download is HERE.

2 Description

2.1 NETS

NETS data are all in the wide data structure, i.e. existing one variable - *Dunsnumber* - that uniquely identifies each row of observation. The rest variables in each NETS subdata record the corresponding latest info for this establishment.

To construct a long dataset for all establishments in NETS, the subdata **Emp** - Employment - is used to determine the lifetime of each establishment. The data **Emp** contains the establishment employment information from 1990 to 2020. An establishment is identified as active only if the employment in that year is nonzero. Otherwise, the establishment is defined as "dead". The data **Emp** is reshaped from wide to long and reduced to two variables only: *Dunsnumber* and *Year*, where variable *Year* shows an active year of a certain establishment. The resulting data is defined as **Est_Active_Year**

The second step is to connect each establishment with its own geographic information - Latitude, Longitude, Address, City, State, and Zipcode - and the name of the establishment onto **Est_Active_Year**. The Latitude and Longitude information can be found in another subdata **Misc** while establishment name, address, and others are stored in subdata **Company**. The merge uses variable *Dunsnumber* as identifier. The resulting data is defined as **Est_ll_address**

The following step is cleaning the data and keeping only correctly recorded establishments. The filtering conditions are

- $Latitude \in [25, 49]$
- $Longitude \in [-125, -67]$

To be noted, the variable *Longitude* is originally stored as a positive float number in data. It is necessary to change it into a negative number for a proper longitude format. More importantly, NETS does not always have an accurate geographic code for each establishment. It might not find the exact latitude and longitude of the establishment. Therefore, for accuracy, resulting data is restricted to only accurately located establishments by forcing variable *Levelcode* = "D". The documentation of variable *Levelcode* can be found by clicking HERE. The resulting data is defined as **Est_ll_address_clean**.

Based on the cleaned data **Est_ll_address_clean**, two data sets are created

- **Net_ll**: Latitude-Longitude data
  - *Dunsnumber*
  - *Year*
  - *Latitude*
  - *Longitude*
- **Net_address**: Name-Address data
  - *Dunsnumber*
  - *Year*

- *Name_net* (Establishment name)
- *Street_net* (address)
- *City_net*
- *State_net*
- *Zip_net*

2.2 TRI

The TRI is a long data recording at Chemical-TRI facility-Year level, i.e. each observation reports the amount of a specific chemical released by a given TRI facility in a certain year. Each TRI facility is assigned one unique identification number, *Trifd*.

To build a long table for the geographic information of TRI facilities, the original TRI data is now compressed to TRI facility-Year observation for unique year-entity geographic records. The resulting data contains the following set of variables

- *Trifd*: uniquely identifying each TRI facility
- *Reportingyear*: Year variable
- *Facilityname*: name of the TRI facility
- *Facilitystreet*: address of the TRI facility
- *Facilitycity*: city where the TRI facility locates
- *FacilityState*: state where the TRI facility locates
- *Facilityzipcode*: zipcode of the TRI facility
- *Latitude*: latitude of the TRI facility
- *Longitude*: longitude of the TRI facility

The resulting data is defined as **Tri_ll_address**.

The TRI geographic data is cleaned in the following order:

- Drop those missing records in variable *Trifd*
- If missing either *Latitude* or *Longitude*, replace both variables with the mode of each within the same TRI facility, i.e. *Trifd*
- Drop observations still with missing *Latitude* or *Longitude*
- Restrict *Latitude* $\in [25, 49]$ and *Longitude* $\in [-125, -67]$

The second step of the cleaning process assumes the TRI facility identification number does not change with a facility move and TRI facility does not move. The resulting data is defined as **Tri_ll_address_clean**.

Similar to NETS, two data sets are constructed from cleaned data **Tri_ll_address_clean**

- **Tri_ll**:
  - *Trifd*
  - *Year*
  - *Latitude*
  - *Longitude*
- **Tri_address**:
  - *Dunsnumber*
  - *Year*
  - *Name_tri* (Facility name)
  - *Street_tri*
  - *City_tri*
  - *State_tri*
  - *Zip_tri*

3 Nearest Establishments

The whole process is conducted at the Year level to reduce the computation pressure. The following steps describe the flow chart within one certain year, e.g. *Year* = 2018 for illustration.

The first match step is to utilize Geocode Package "**Geonear**" to find all NETS establishments within 500 meters of each TRI facility, using *Latitude* and *Longitude* to identify the exact geo-location for NETS establishments and TRI facilities. For year 2018, the match result returns 1,040,966 total observations where each is a unique combination of TRI facility and NETS establishment satisfying the within-500-meter condition for 21,866 reported TRI facilities. On average, one TRI facility could find **48** NETS establishments within a 500-meter radius. The match result also reports the distance between a TRI facility and a corresponding NETS establishment.

To prepare for the matching process, several new variables are constructed. All new and original variables used in match are as follows:

- *Name_tri*: original name of a TRI facility
- *Name_tri_prune*: name of a TRI facility stripped off space and any non-alphabet characters (e.g. & ' -)
- *Name_tri_f10*: first ten characters of original name of a TRI facility
- *Name_tri_f4*: the first four characters of *Name_tri_prune*
- *Name_tri_f8*: the first eight characters of *Name_tri_prune*
- *Street_tri*: original street of a TRI facility
- *Street_tri_prune*: street of a TRI facility stripped off any non-alphabet characters
- *Name_net*: original name of a NETS establishment
- *Name_net_prune*: name of a NETS establishment stripped off space and any non-alphabet characters

- *Name_net_f10*: first ten characters of original name of a NETS establishment
- *Name_net_f4*: the first four characters of *Name_net_prune*
- *Name_net_f8*: the first eight characters of *Name_net_prune*
- *Street_net*: original street of a NETS establishment
- *Street_net_prune*: street of a NETS establishment stripped off any non-alphabet characters

4 Match Process

4.1 Preface

Among those establishments that have the 500m or less distance to a given TRI facility, the names of those establishments are fuzzy compared with the name of the given TRI facility using **Matchit** package. The similarity method for the fuzzy match contains *token*, *firstgram, n*, *ngram, n*, and plus *soundex*.

Similarity method *token* is defined as comparing the similarity of each phrase between two names. For instance, if one name is "ABC CO" and the other is "ABC INC", the algorithm disassembles into two phrases - "ABC" and "CO" - for the first and two phrases - "ABC" and "INC" - for the second, and reports 50% similarity. If the second is "ABC INC CO", then reports 81% similarity.

In comparison, *firstgram, n* compares the first n characters for each phrase, such that, if $n = 2$, then the above example will be "AB" and "CO", as well as "AB" and "IN" and "CO".

The more complex method is the last one *ngram, n*. While *firstgram, n* only takes the first n characters in each phrase, *ngram, n* constructs a set of n-length phrases using iteration. For example, the name "ABC INC" under *ngram, 3* will be "ABC", "BC ", "C I", " IN", and "INC". As a result, this method gives only 44% similarity for "ABC INC" and "ABC CO". 71% of similarity if "ABC INC CO" and "ABC CO".

Lastly, the *soundex* method is a Phonetic algorithms. It recodes text using the Soundex algorithm (e.g. "John Smith" recodes to J525). The Soundex code consists of a letter followed by three numbers: the letter is the first letter of the name and the numbers encode the remaining consonants. Similar-sounding consonants are encoded by the same number. Results would be only either 0 or 1.

4.2 Match #1

Based on this algorithm, the correctly corresponding establishment among close ones should have the same phrases inside the name as TRI facility's name. Define that, for TRI facility i and a NETS establishment j

- $distance_{i,j}$ is the distance between TRI facility i and NETS establishment j
- $distanec_{\min,i}$ is the minimum distance between a NETS establishment to the TRI facility i
- $sim\_name_{i,j}$ is the original name fuzzy match *token* similarity between TRI facility i and NETS establishment j

The qualifications for the "first-best" match should satisfy either one of the four conditions below

$$\mathbb{1}(distance_{i,j} = distance_{\min,i}) \quad \& \quad sim\_name_{i,j} > 0.5$$

$$\mathbb{1}(distance_{i,j} = distance_{\min,i}) \quad \& \quad 0.5 \geq sim\_name_{i,j} > 0 \quad \& \quad \mathbb{1}(Street\_tri_i = Street\_net_j)$$

$$\mathbb{1}(distance_{i,j} \neq distance_{\min,i}) \quad \& \quad sim\_name_{i,j} > 0.6$$

$$\mathbb{1}(distance_{i,j} \neq distance_{\min,i}) \quad \& \quad 0.6 \geq sim\_name_{i,j} \geq 0.5 \quad \& \quad \mathbb{1}(Street\_tri_i = Street\_net_j)$$

The idea of the first match qualification conditions is simple

- 1. among the closest ones, identify an establishment as matched if whose name has at least half of the phrases appearing in the corresponding TRI facility's name
- 2. among the closest ones, identify an establishment as matched if 1) whose address is identical to a TRI facility's address and 2) whose name is not entirely different from a TRI facility's name
- 3. for those not closest ones, identify an establishment as matched if whose name is highly similar to a TRI facility's name (usually similarity above 0.6)
- 4. for those not closest ones, identify an establishment as matched if 1) whose name is similar to a TRI facility's name (greater than 0.5 to avoid similar suffix) and 2) whose address is identical to a TRI facility's address

The qualified results are defined as **Match_1**

4.3 Match #2

Among those above unmatched, the second match process is to look at a more restricted name format. Define that

- $sim\_name\_f10\_f5_{i,j}$ is the first-ten-character name fuzzy match similarity between TRI facility i and NETS establishment j under *firstgram, 5* method.

The qualification of the "second-best" match is

$$sim\_name\_f10\_f5_{i,j} > 0.5$$

This idea follows that both a TRI facility and a NETS establishment should place the most distinguishable or unique phrase at the beginning of a name. If the initial parts of the two names are similar, then the NETS establishment is identified as matched. The qualified results are defined as **Match_2**.

4.4 Match #3

After the first two filters, the leftover suffers more complex problems, and thus, needs more constraints from various aspects to extract "match" ones. Define that

- $sim\_f8\_n4_{i,j}$ is the *ngram, 4* similarity for *Name_tri_f8* and *Name_net_f8*
- $sim\_street\_t_{i,j}$ is the *token* similarity of *Street_net* and *Street_tri*
- $sim\_street\_pt_{i,j}$ is the *token* similarity of *Street_net_prune* and *Street_tri_prune*

The "third-best" conditions are as follows

$$\mathbb{1}(Name\_net\_f4 = Name\_tri\_f4) \quad \& \quad sim\_f8\_n4_{i,j} > 0.5$$

$$\mathbb{1}(Name\_net\_f4 = Name\_tri\_f4) \quad \& \quad 0.5 \geq sim\_f8\_n4_{i,j} > 0.2 \quad \& \quad sim\_street\_pt_{i,j} > 0$$

$$\mathbb{1}(Name\_net\_f4 = Name\_tri\_f4) \quad \& \quad sim\_f8\_n4_{i,j} = 0.2 \quad \& \quad sim\_street\_t_{i,j} > 0.5$$

$$\mathbb{1}(Name\_net\_f4 \neq Name\_tri\_f4) \quad \& \quad sim\_f8\_n4_{i,j} > 0.2 \quad \& \quad sim\_street\_pt_{i,j} > 0.6$$

These conditions are more applied to special cases rather than general ones. And the threshold / cut-off rate is more of a tradeoff between the false positive and false negative, which is difficult to illustrate in readable examples. The target of each qualification is as follows

- 1. Similar to the unique-words-starts-first case, this condition deals with initial phrases suffering from special characters. E.g. "O'nell" and "Onell". The length of the phrase is less than the previous so that it could avoid potential different-suffix disturbance

- 2. and 3. are amendments to condition 1. to utilize address similarity if name similarity suffers from suffix disturbance
- 4. This is to identify a highly similar address and similar name case. For instance, a NETS establishment "SEMICNDCTOR CMPONENTS INDS LLC" with address "23400 NE GLISAN ST" corresponds to "ONSEMI" with address "23400 NORTHEAST GLISAN ST"

5 Result

Results are shown in table 15

Year	# of TRI Facilities	# of Match	Match Ratio
1990	25152	13199	0.5247694
1991	24836	13057	0.52572878
1992	24572	13001	0.52909816
1993	24101	13186	0.54711423
1994	23319	12773	0.54775076
1995	22777	12655	0.55560434
1996	22422	12577	0.56092231
1997	22246	12692	0.57052953
1998	24057	13624	0.56632165
1999	23189	13204	0.56940791
2000	23955	13721	0.5727823
2001	25631	14848	0.57929851
2002	24982	14812	0.59290689
2003	24444	14604	0.59744723
2004	24230	14491	0.59806026
2005	24064	14439	0.60002493
2006	23627	14329	0.60646718
2007	23119	14025	0.60664389
2008	22562	13802	0.61173655
2009	21696	13419	0.61850111
2010	21588	13467	0.62381879
2011	21703	13755	0.63378335
2012	21892	14105	0.64429929
2013	22088	14309	0.64781782
2014	22173	14476	0.6528661
2015	22187	14501	0.65358093
2016	22032	14813	0.67234023
2017	21861	14775	0.67586112
2018	21866	14861	0.67963962
2019	21667	14786	0.68242027
2020	21090	14499	0.68748222

Table 1: Match Results

6 Classification Method

This section is an extension of the previous result. Denoting match result sent from Prof. Miguel as **"match_outside"**, results from checking the match from both **"match_outside"** and **"match_my_algorithm"** report only half of both resulting data are matched. In specific, given the most relaxed constraints, the match between these two data is exhibited in Table 2

Results	# of obs
Not Matched	626383
From **"match_outside"**	311377
From **"match_my_algorithm"**	315006
Matched	306981

Table 2: Checking Match

Therefore, based on the match checking, the classification tree method is used to improve finding the right combination of NETS establishment and TRI facility. The following sub-sections will demonstrate

- 1. Construction of training and testing samples;
- 2. Classification algorithms utilized;
- 3. Results on applying classification methods.

6.1 Construction of Training and Testing Samples

The training and testing samples are constructed from three sets of data:

- **All within-500-meter combinations**,
- **Matched_my_algorithm**
- **Match_outside**

Denoted the overlapping combination from **Matched_my_algorithm** and **Match_outside** as the identified "true" matched combination, while the identified "true" unmatched comes from any "**All within-500-meter combinations**" combination neither in **Matched_my_algorithm** nor **Match_outside**. Here is a summary Table 3 of the three datasets and the constructed training-testing sample **Train-Test** Be noted, the change

Data	# of obs	Unit of obs	Match/Unmatch Ratio
All within-500-meter combinations	32298318	TRI-NETS-Year	1 / 98 $\approx$ 0.2
Matched_my_algorithm	618358	TRI-NETS-Year	
Match_outside	621987	TRI-NETS-Year	
Train-Test	620043	TRI-NETS-Year	1 / 1.05 $\approx$ 1

Table 3: Caption

from *Match / Unmatch Ratio* ≈ 0.2 in **All within-500-meter combinations** to approximate 1 in **Train-Test** is due to balancing the match and unmatch observations. The balancing method is to randomly drop off unmatch observation upto *Match / Unmatch Ratio* is almost 1.

6.2 Classification Algorithms Utilized

There are two classification methods utilized (three are tested, top two are chosen): Cubist Tree and Logistic Regression. The data **Train-Test** is split by 70%/30% into the Training sample and Testing sample, respectively. To use the classification method, the following set of features is generated, similar to those in the previous attempt (difference #1: for street info, changed all directional abbreviation into full name, like "N" into "NORTH" and "SE" into "SOUTHEAST")

- *sim_name_t*: name similarity under *token*
- *sim_name_f4*: name similarity under *firstgram, 4*
- *sim_name_s*: name similarity under *soundex* – pronunciation
- *sim_name_prune_substr_n4*: first-10-character-and-w/o-space name similarity under *ngram, 4*
- *sim_name_f2t_n3*: first two token similarity under *ngram, 3*
- *idx_first5char*: identifier for the same first 5 characters in name
- *sim_street_t*: street similarity under *token*
- *sim_street_s*: street similarity under *soundex*
- *street_num_idx*: identifier for the same street number

Performing the trained model upon the testing sample, the predicted results are summarized in Table 4

Model	Out-Of-Sample R^2	Confusion Matrix			
				Labeled	
				1	0
Cubist Tree	0.9978	Predict	1	92279	102
			0	5	94431
				Labeled	
				1	0
Logistic Regression	0.9947	Predict	1	92224	186
			0	60	94347

Note: 1 denotes matched combination; otherwise, 0 is not matched combination

Table 4: Classification Models Performance

6.3 Results on Applying Classification Methods

Back to **All within-500-meter combinations**, both Cubist Tree and Logistic Regression were applied on this full sample. Observations predicted as matched under either Cubist or Logistic are identified as a matched combination of TRI-NETS at a given year. The classification results, denoted as **"match_classification"**, are as follows: out of 32,298,318 TRI-NETS-Year observations with 5,306,698 unique TRI-NETS combinations, 683,981 TRI-NETS-Year observations are identified as matched observations. Cross-check with **Match_outside** at TRI-NETS-Year level, results are shown below in Table 5

Thus, the two-classification results give a slightly more identified match than the previous attempt (my manual algorithms). With a manual check into the **"match_classification"** - **"match_outside"** matched results, 49 out of the first 1000 $\approx$ 5.0% of the observations are either a false positive or ambiguous/uncertain. Around one-fifth are ambiguous/uncertain while the rest four-fifths are likely to be false positive (even though they have identical street information).

Results	# of obs	# of unique TRI	# of unique TRI-NET comb
Not Matched	665128	47934	88188
From **"match_outside"**	301567	34633	34706
From **"match_classification"**	363561	26989	56968
Matched	320420	27038	27080

Table 5: Cross Checking

7 Clean-Up

7.1 Recap

All results are built upon three datasets

- **National Establishment Time Series (NETS)**: original Establishment-Year-level data provided by Walls & Associates
- **Toxic Release Inventory (TRI)**: original Facility-Year-Chemical-level data provided by Environmental Protection Agency
- **NETS-Match**: an external match between TRI facilities to NETS establishments by Walls & Associates

Here is a table of comparison for the three datasets.

Data	Unit of Analysis	Year Range	# of Obs
NETS	Establishment-Year	1990 to 2020	602,948,028
TRI	Facility-Chemical-Year	1990 to 2020	2,615,683
NETS-Match	Establishment-Facility-Year	1987 to 2019	596,120

Table 6: Caption

7.1.1 Prepare data

From the first dataset **NETS**, we construct the following 3 sets of data for use

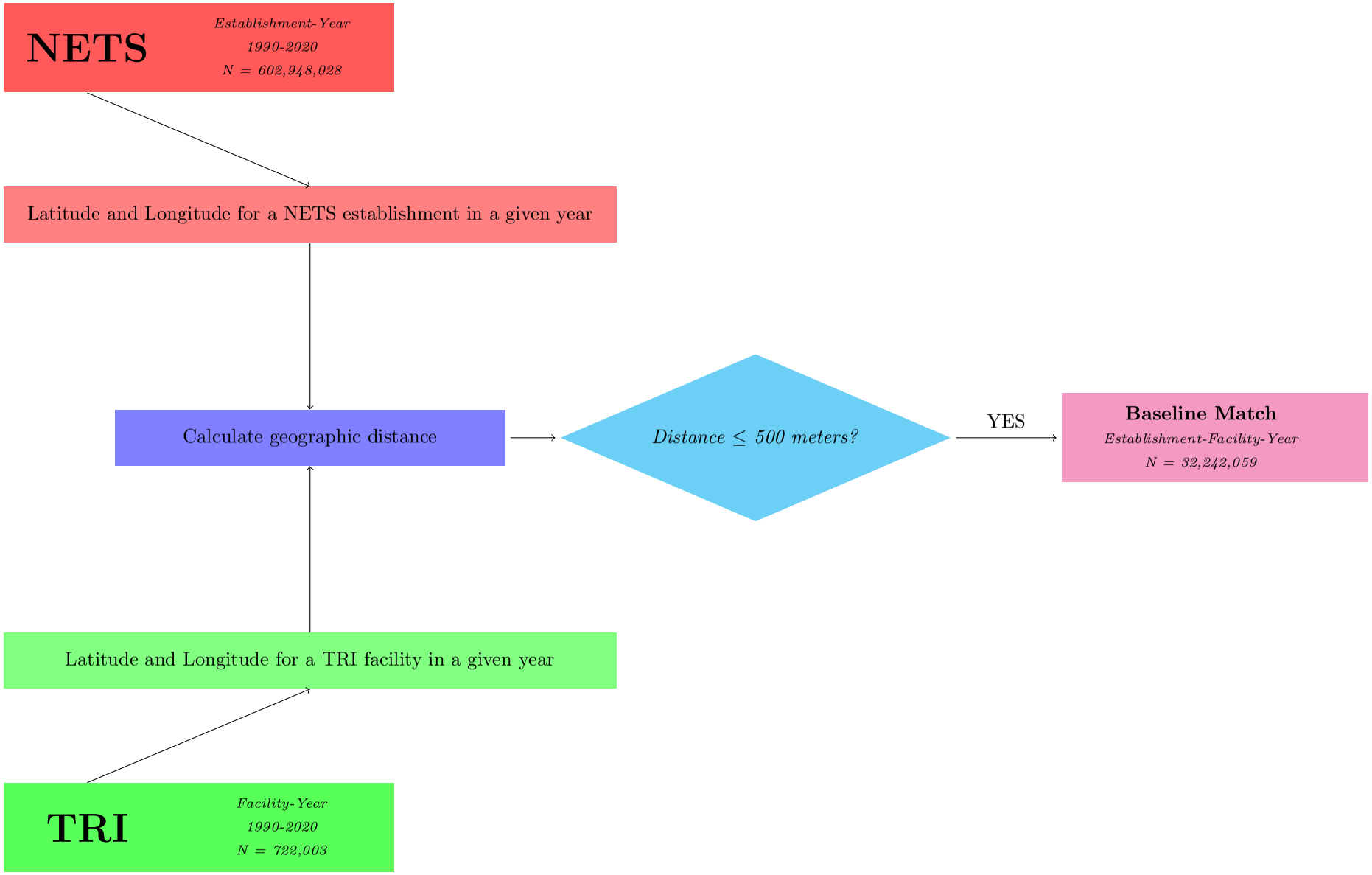
- **NETS-Latitude & Longitude**: contains each establishment's latitude and longitude information in a given year
- **NETS-Address**: contains each establishment's name, address, city, state, and zip code information in a given year
- **NETS-Naics**: contains each establishment's NAICS Code in a given year

From the second dataset **TRI**, we obtain similar 3 sets of data

- **TRI-Latitude & Longitude**: each facility's latitude and longitude information at a given year
- **TRI-Address**: each facility's name, address, city, state, and zip code information at a given year
- **TRI-NaicsList**: contains all NAICS Code that a TRI facility involves in a given year

7.1.2 Construct a match baseline data

Using **NETS-Latitude & Longitude** and **TRI-Latitude & Longitude**, the match baseline data is constructed from matching all possible combinations between a NETS establishment and a TRI facility as long as the geographic distance between a facility and an establishment is within 500 meters.


NETS
Establishment-Year
1990-2020
N = 602,948,028
Latitude and Longitude for a NETS establishment in a given year
Calculate geographic distance
Distance ≤ 500 meters?
YES
Baseline Match
Establishment-Facility-Year
N = 32,242,059
Latitude and Longitude for a TRI facility in a given year
TRI
Facility-Year
1990-2020
N = 722,003

Then, upon the baseline match data, the two previous attempts–Manual Algorithm and Classification Tree–are used to generate corresponding matching results. Thus, denote the matched result from the first and second attempts as, **ATT1** and **ATT2**, respectively.

7.1.3 Glance on NETS-Match

Before going into finding the correct NETS establishment - TRI facility matches, some statistical descriptions are exhibited to give a brief summary of the compositions of **NETS-Match**. The data **NETS-Match** contains five variables:

- *Trifd*: unique identifier for each TRI facility
- *Dunsnumber*: unique identifier to each NETS establishment
- *Firstreport*: the year PRIOR to when a NETS establishment is considered to be active
- *Lastreport*: the last year when a NETS establishment is active
- *Year*: the year when the match is considered to exist

Thus, each row records a NETS establishment - TRI facility match in a given year. However, when comparing the TRI facilities in **NETS-Match** to those in **TRI**, there exist some unmatched TRI facilities. Around one TRI facility per year shown in **NETS-Match** never appears in **TRI**. Details on unmatched TRI facilities are exhibited in Appendix A.1.

Second to notice, **NETS-Match** reports a match belonging to either of these three cases

- Mostly, there is a "one-NETS-establishment"-to-"one-TRI-facility" matches (around 98%)
- Some are "one-NETS-establishment"-to-"multiple-TRI-facilities" matches (around 1.5%)
- Rare are "multiple-NETS-establishments"-to-"one-TRI-facility" matches (around 0.5%)

On average, however, **NETS-Match** identifies 1% more TRI facilities than NETS establishments every year. In the whole sample, the number of unique TRI facilities to the number of unique NETS establishments is

$$\frac{\#\text{ of Total TRI facilities identified}}{\#\text{ of Total NETS establishments identified}} = 1.04$$

Excluding those unmatched records in **NETS-Match**, here is a statistical summary of the distance between matches:

	Mean	Std	Min	P1	P25	Median	P75	P99	Max
Distance	32.80675	233.8422	0	.0050444	.0515921	.156498	.7065313	1098.767	12628.21

Measured in unit of kilometers. The distance summary in this table rationales the later constrain of "distance$\leq 10 miles$"

Table 7: Distance Summary

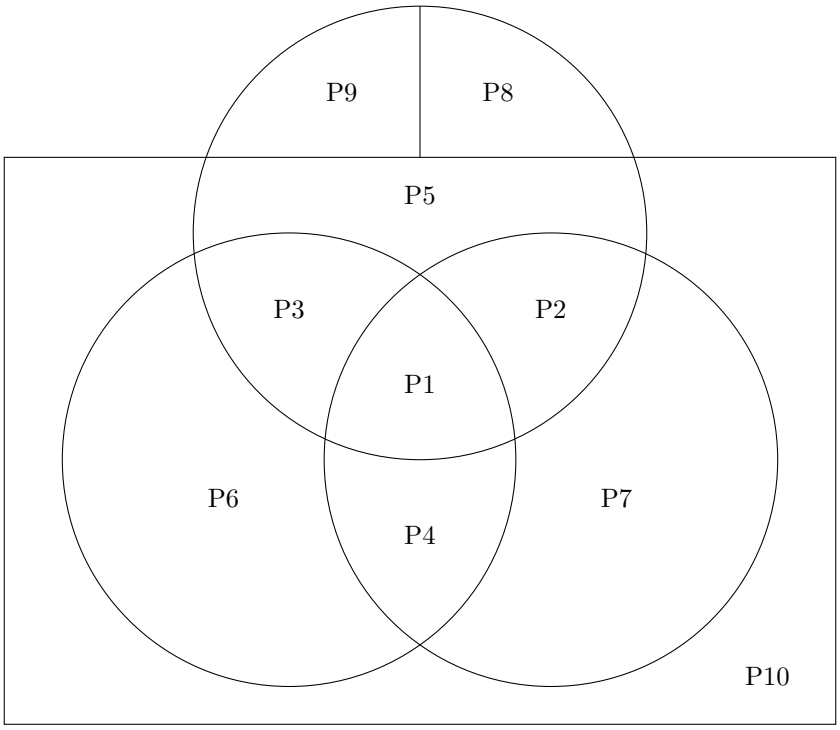
Lastly, the variable Year in the original data ranges from 1987 to 2019. It is now pruned down to 1990 to 2019 to overlap with **TRI** and **NETS**.

7.2 All potential combinations

To better distinguish each resulting data from the previous two attempts and the external source, remind that

- **NETS-Match**: a NETS-TRI matches data provided by Walls & Associates
- **ATT1**: matches from Attempt 1 as for Manual Algorithm 4.
- **ATT2**: matches from Attempt 2 as for Classification Tree methods 6

The relationship between these three sets of data can be summarized in the following graph,



$$P1 + P2 + P3 + P5 + P8 + P9 \longrightarrow \textbf{NETS-Match}$$

$$P1 + P2 + P4 + P7 \longrightarrow \textbf{ATT1}$$

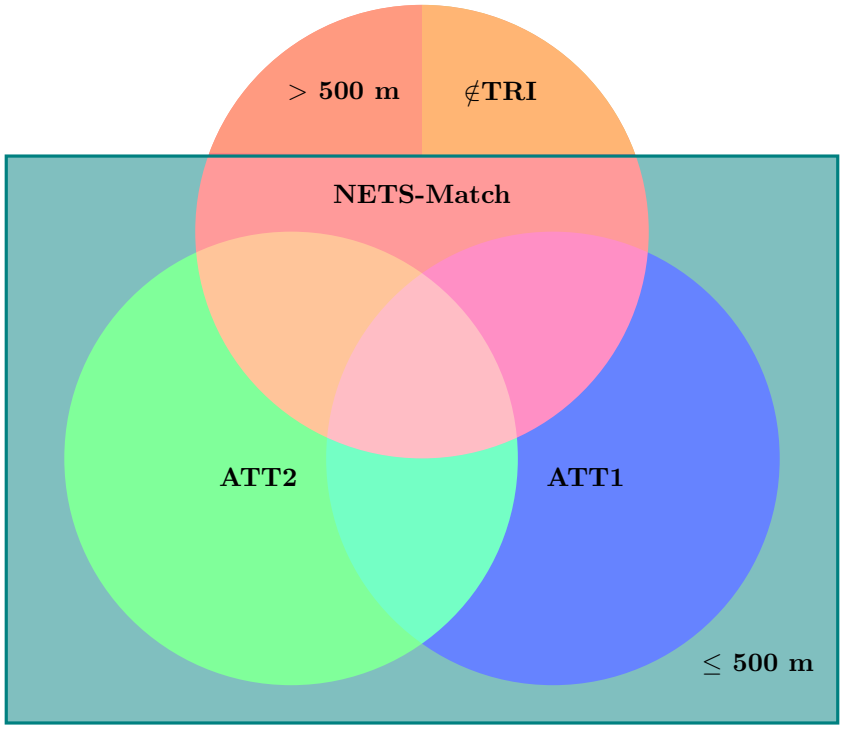
$$P1 + P3 + P4 + P6 \longrightarrow \textbf{ATT2}$$

$$P1 + P2 + P3 + P4 + P5 + P6 + P7 + P10 \longrightarrow \textbf{Baseline Match}$$

In specific, each part $P*$ represents:

- $P1$–interception of **NETS-Match**, **ATT1**, and **ATT2**
  (i.e. **NETS-Match** $\cap$ **ATT1**$\cap$**ATT2**)
- $P2$–interception of **NETS-Match** and **ATT1**, but not in **ATT2**
  (i.e. **NETS-Match** $\cap$ **ATT1** $\cap$ **ATT2**^{c})
- $P3$–interception of **NETS-Match** and **ATT2**, but not in **ATT1**
  (i.e. **NETS-Match** $\cap$ **ATT1**^{c} $\cap$ **ATT2**)
- $P4$–interception of **ATT1** and **ATT2**, but not in **NETS-Match**
  (i.e. **NETS-Match**^{c} $\cap$ **ATT1** $\cap$ **ATT2**)
- $P5$–**NETS-Match** that neither in **ATT1** nor **ATT2**, and match distance is within 500 meters
- $P6$–**ATT2** that neither in **NETS-Match** nor **ATT1**
- $P7$–**ATT1** that neither in **NETS-Match** nor **ATT2**
- $P8$–**NETS-Match** that has TRI facility-Year not in **TRI**
- $P9$–**NETS-Match** has overlap TRI facility-Year in **TRI** but match distance is larger than 500 meters
- $P10$–TRI facility-NETS establishment matches whose distance is within 500 meters, but identified as false matches

The details of the number of observations and proportions to the whole set as shown in graph 7.2 are summarized in Appendix B. Overall, the relationship is simplified into the following, where the most important/distinguishable feature is highlighted



Therefore, based on the interceptions and unions, define 5 Groups of the dataset as follow

- Group 1 – all from **NETS-Match**, excluding ∉ **TRI**
- Group 2 – ∼ **NETS-Match** & **ATT1** & **ATT2**
- Group 3 – ∼ **NETS-Match** & ∼ **ATT1** & **ATT2**
- Group 4 – ∼ **NETS-Match** & **ATT1** & ∼ **ATT2**
- Group 5 – ∼ **NETS-Match** & ∼ **ATT1** & ∼ **ATT2**

Table 8 summarizes the number of matches in each Group. Table 14 in Appendix B.2 includes all groups.

Group	# of Obs	Percent
Group 1	547563	0.585
Group 2	311640	0.338
Group 3	60080	0.065
Group 4	1594	0.002
Total	920877	

Table 8: Descriptive Summary (Ex. Group 5)

And the cleanup is conducted on top of Group 1 to Group 4. Denote this Group 1 to Group 4 set as **Match-Candidates**.

7.3 Filtering Conditions for Each Group

7.3.1 Append Naics Codes

First of all, for each match in **Match-Candidate**, the establishment's NAICS Code and the list of NAICS Code related to the facility are appended. In NETS, only one NAICS Code is recorded for one NETS establishment in a given year; meanwhile, TRI keeps track of all NAICS Code involved in one TRI facility in a given year.

7.3.2 Filter

Thus, each match contains

- Group Identifier: Identifies the match is from which Group (1/2/3/4)
- Year
- Trifd
- TRI-Address related:
  - Address
  - City
  - State
  - Zipcode
- TRI-Naics Code: All Naics Code involved
- Dunsnumber
- NETS-Address related:
  - Address
  - City
  - State
  - Zipcode
- NETS-Naics Code

The filtering conditions for each group are simple and straightforward.

$$\text{Group 1}: \begin{cases} \text{Distance} \leq 10 \text{ Miles} \\ \text{State (TRI)} = \text{State (NETS)} \end{cases}$$

$$\text{Group 2 or 3}: \begin{cases} \text{Naics Code-Sector \& Subsector (TRI)} = \text{Naics Code-Sector \& Subsector (NETS)} \\ \text{Naics Code-Industry Group (TRI) - Naics Code-Industry Group (NETS)} \leq 6 \text{ Industry Groups} \\ \text{Distance} \leq 10 \text{ Miles} \\ \text{State (TRI)} = \text{State (NETS)} \end{cases}$$

$$\text{Group 4}: \begin{cases} \text{Address (TRI)} = \text{Address (NETS)} \\ OR \\ \text{Naics Code (TRI)} = \text{Naics Code (NETS)} \end{cases}$$

Table 9 shows the change of the numbers of obs in each group before and after filtering.

Group Index	Before Filer	After Filter	Change	Change (Percent)
1	547563	518311	29252	5.3%
2	311640	105122	206518	66.3%
3	60080	13118	46962	78.2%
4	1594	31	1563	98.1%

Table 9: Filter Result

To better explain the second filtering condition for "Group 2 or 3",

Naics Code-Industry Group (TRI) - Naics Code-Industry Group (NETS) $\leq$ 6 Industry Groups

let us take two examples:

- Example 1:

$$\begin{cases} \text{NAICS Code (TRI)} = 332111 - - \textit{Iron and Steel Forging} \\ \text{NAICS Code (NETS)} = 332613 - - \textit{Spring Manufacturing} \end{cases}$$

  Therefore, the second condition says it is a qualified one because the difference between these two is 502, within the 6 Industry Groups (i.e. within 600).

- Example 2:

$$\begin{cases} \text{NAICS Code (TRI)} = 311111 - - \textit{Dog and Cat Food Manufacturing} \\ \text{NAICS Code (NETS)} = 311941 - - \textit{Mayonnaise, Dressing, and Other Prepared Sauce Manufacturing} \end{cases}$$

  Although these two NAICS Codes fall into the same industry, a mayonnaise product is clearly not pet food. And thus, these match is unqualified.

Despite one could argue it is less reliable to use the difference in categorical variables as a cardinal indicator, NAICS Code still can indicate similar activities within the same industrial sector. Criteria for comparing differences between NAICS Code could change to analyze its sensitivity as in Appendix C.1.

8 Matching Capital IQ (new)

8.1 Preparing Capital IQ

The Capital IQ–Computstats data is directly downloaded from the WRDS website. This dataset contains a large list of U.S. and Canadian publicly-traded corporations' information on their fundamental and market activities. In this task, we focus on only annual fundamental data within North American Companies, ranging date from 1990 to 2020. The link to the data download website can be accessed here Capital IQ.

8.2 Processing Capital IQ

Downloaded on Nov 14, 2022, denote Capital IQ data as **CIQ**. Since the target of this task is to match the parent companies from NETS-TRI data to companies in **CIQ**, corporation names, denoted as *conm*, in **CIQ** is crucial for name fuzzy match. The address information is less important here for the following two reasons:

- **TRI** only records parent company names
- **NETS** only records headquarter / subsidiary geographic information, rather than the parent company's

Therefore, it leaves us only the fuzzy match of names between two datasets.

8.2.1 Capital IQ Filtering

First of all, there are some missing values in NAICS Code / SIC Code in **CIQ**. I manually searched these firms on the Internet and filled them in. Then, since the goal of this research question concentrates on most manufacturing industries, companies from the Finance Sector are dropped out by dropping those whose SIC Code is between 6000 to 6999 (representing Finance or Banking). Thus, the cleaned **CIQ** has the following set of variables

- *gvkey*: Unique company identifier
- *fyear*: Fiscal Year variable
- *datadate*: The exact data time when the annual report is collected
- *conm*: Company listed name

The data contains 20,445 unique companies throughout the sample, ranging from 1990 to 2020. On average, there are 7,154 corporations each fiscal year.

8.3 Name Fuzzy Match: TRI-CIQ

Since the original company names are not as messy as the previous TRI-NET ones, there is not too much cleaning conducted on the name. The entire name cleaning can be summarized into the following three steps:

- Drop "CL -A", "CL -B" and "OLD"
  - "CL -A": Share Class A
  - "CL -B": Share Class B
  - "OLD": Representing old capital structure of the same company
- Drop symbols
  - No alphabetic symbols, such as , . # &, etc
- Drop common suffix
  - Common suffix that shows up at the very end of the name, including "CO INC CORP GROUP INDS LTD LP LLC PLC GP GRP CP"

These steps are conducted identically to both **TRI** and **CIQ**. It potentially results in that one cleaned name could match multiple uncleaned names. However, one could uniquely link back by using the company lifetime range (from the beginning year of a company to the end of it). For example, there exist two rows of observation: [*gvkey*, *conm*]

- 031625, CARBON ENERGY CORP - OLD
- 186558, CARBON ENERGY CORP

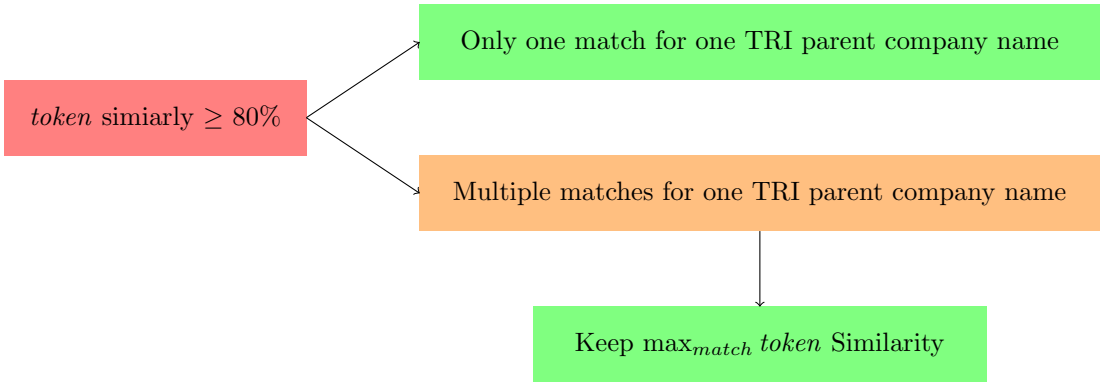
both have the clean name as "CARBON ENERGY". But the first one spans life from 1999 to 2002, while the second is from 2010 to 2019.

8.4 Fuzzy Match Criteria

There is only one method used in Fuzzy Match: *token*. Both **TRI** and **CIQ** have unique company names not as many as facility names/establishment names in **TRI-NETS**, and thus, only one matching method is used. Remind that *token* is to check the frequency of each word in both names Details and examples can be found in the previous fuzzy match 4.1. Because the *token* fuzzy match returns matches which at least have one common word, it greatly reduces those false matches and only reports a result of 369,182 potential matches, with the least *token* similarity at 50% (up to 95 percentile).

There have two match criteria:

- First, keep if *token* similarity $\geq 80\%$
- Second, keep only the maximum-*token*-similarity matches if there are multiple matches for one TRI parent company name



Therefore, we have each TRI facility match to one CIQ Company (99% of the cases). There exist around 10 cases where the CIQ Company is acquired by another CIQ Company in a later year. These are solved manually.

8.5 Fuzzy Match Results

Here is a summary table of the match results:

	# of Unique Obs	Total Obs
CIQ	1626	147,044
TRI	9447	147,044
NETS	12730	147,044

Table 10: Unique Observations

Figure 1-4 illustrates the trend of CIQ Companies, TRI Facilities, and NETS Establishments throughout the sample periods. Notice that there is a big drop in 2020. The reason is that there is no data in **NETS-Match** from Walls and Associates, resulting in fewer matches of TRI-NETS in 2020. The data in 2020 could be dropped off to have balanced observations throughout all sample periods.

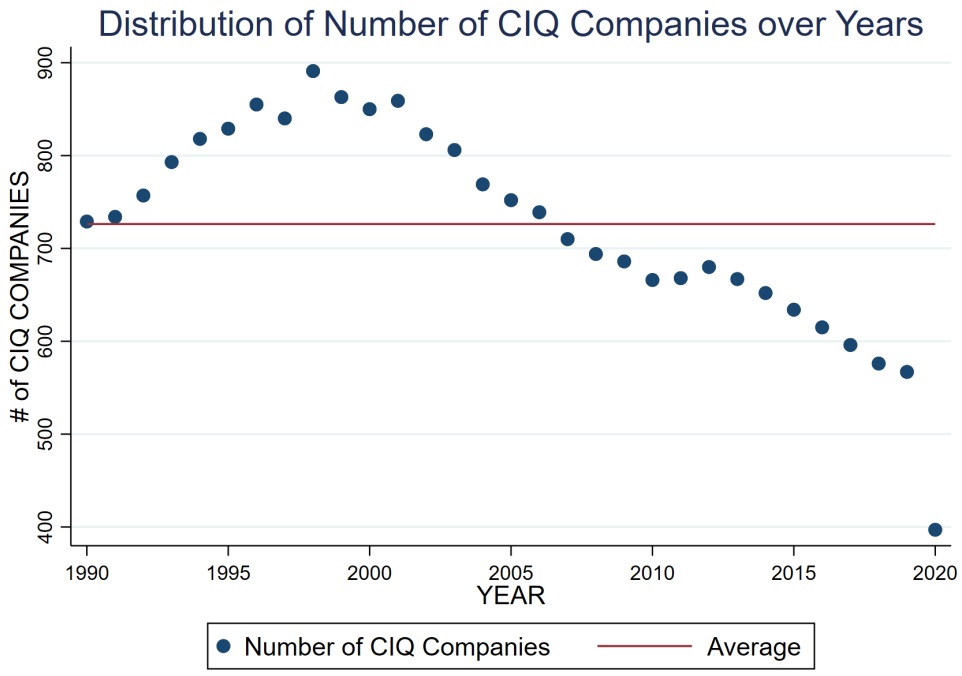


Figure 1: Dist. of CIQ

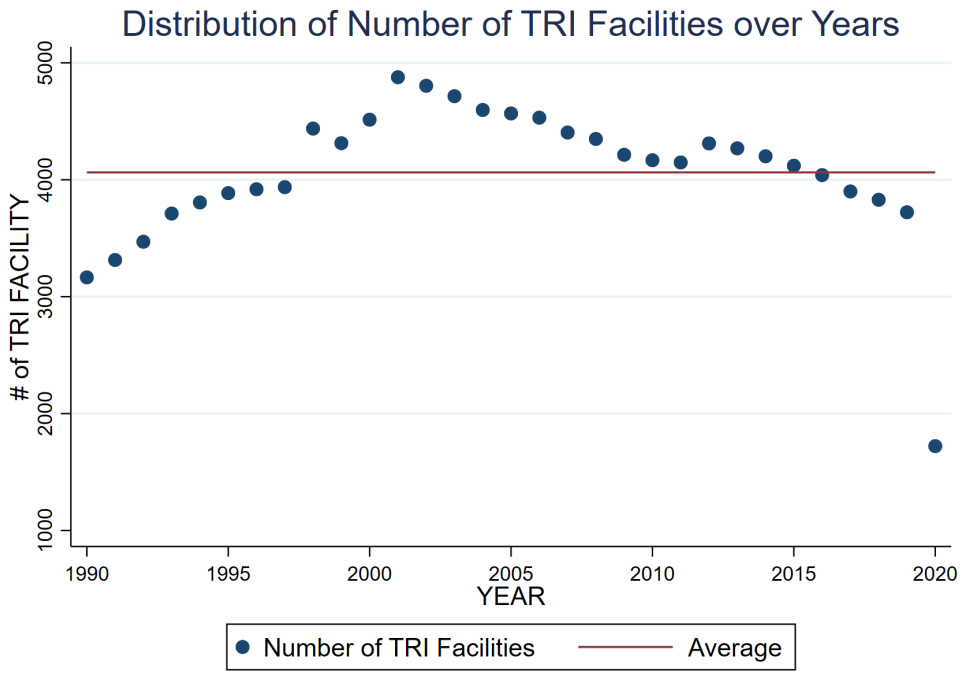


Figure 2: Dist. of TRI

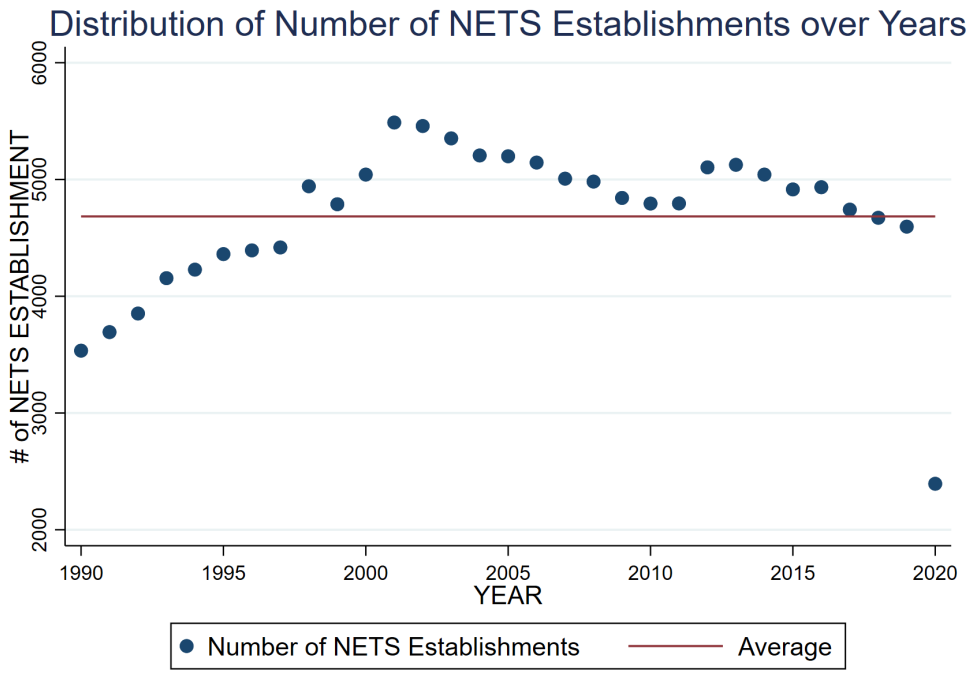


Figure 3: Dist. of NETS

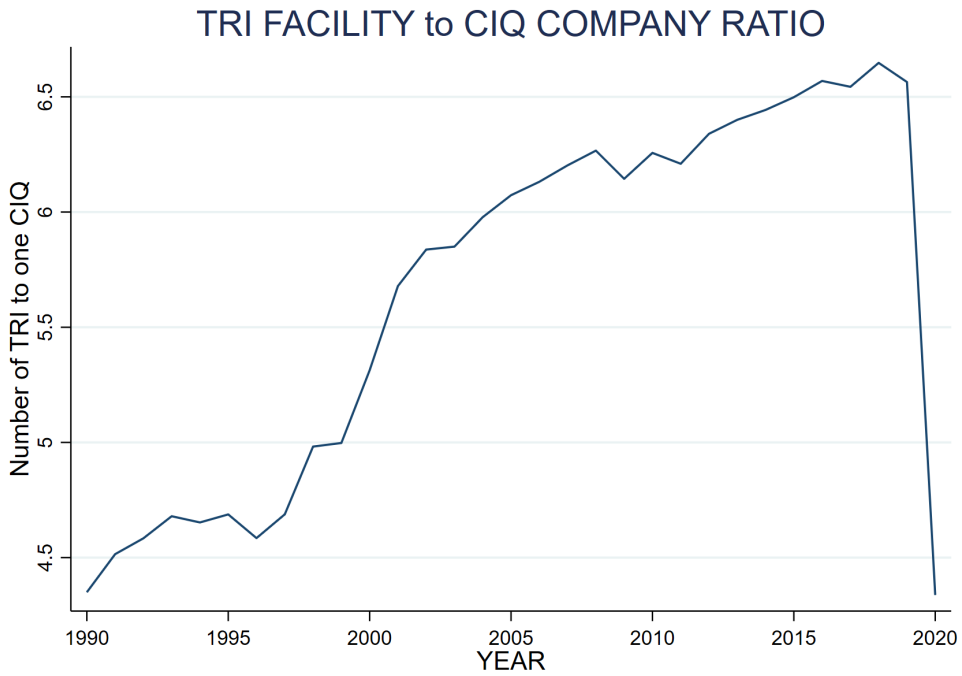


Figure 4: Dist. of TRI to CIQ Ratio

Appendix A NETS-Match

A.1 Unmatch TRI Facilities Distribution

	# of Obs	Percent	Cum.
1990	1284	2.646	2.646
1991	1499	3.089	5.734
1992	1652	3.404	9.138
1993	1850	3.812	12.95
1994	2101	4.329	17.28
1995	2289	4.716	22.00
1996	2449	5.046	27.04
1997	2505	5.162	32.20
1998	2412	4.970	37.17
1999	2663	5.487	42.66
2000	2377	4.898	47.56
2001	1972	4.063	51.62
2002	1956	4.030	55.65
2003	1921	3.958	59.61
2004	1817	3.744	63.35
2005	1721	3.546	66.90
2006	1689	3.480	70.38
2007	1776	3.659	74.04
2008	1759	3.624	77.66
2009	1956	4.030	81.69
2010	1685	3.472	85.17
2011	1446	2.979	88.15
2012	1303	2.685	90.83
2013	1105	2.277	93.11
2014	964	1.986	95.09
2015	797	1.642	96.74
2016	697	1.436	98.17
2017	549	1.131	99.30
2018	327	0.674	99.98
2019	11	0.0227	100
Total	48532	100	100

1. Unmatch Ratio (Unmatch / Total) = 8.15%.
2. Percent denotes unmatched TRI-Year obs each year to total TRI-Year obs

Table 11: Unmatched TRI Facilities over Years

This table contains all TRI facilities that show up in **NETS-Match** but disappear in **TRI**, at a given year. For instance, the year 1990 reports 1284, meaning there are 1284 unique TRI facilities showing up in **NETS-Match** in 1990 but not showing up in **TRI** in the same year.

These TRI facilities could be found in other years in **TRI**. Only less than 15 TRI facilities showing up in **NETS-Match** never appear in **TRI** (out of 57759 unique TRI facilities in **TRI** and 51353 in **NETS-Match**).

Appendix B Composition Summary

B.1 Part Summary

Part (P*)	# of Obs	Percent	Cum.
1	300036	0.923	0.923
2	833	0.00256	0.926
3	20384	0.0627	0.988
4	228670	0.703	1.692
5	67202	0.207	1.898
6	134891	0.415	2.313
7	24127	0.0742	2.388
8	48557	0.149	2.537
9	159108	0.489	3.027
10	3.15e+07	96.97	100
Total	3.25e+07	100	100

Table 12: Part Descriptive Summary

Part (P*)	Mean	Std	Min	P1	P25	Median	P75	P99	Max
1	0.0740	0.0707	0	0.00258	0.0194	0.0497	0.105	0.291	0.311
2	0.0978	0.0810	0.00263	0.00263	0.0305	0.0739	0.149	0.270	0.307
3	0.0855	0.0735	0.000772	0.00357	0.0269	0.0615	0.124	0.296	0.310
4	0.101	0.0865	0	0.00292	0.0277	0.0745	0.160	0.303	0.311
5	49.48	223.5	0	0.00523	0.147	0.754	3.851	1184.2	7846.8
6	0.119	0.0897	0	0.00406	0.0405	0.0972	0.191	0.305	0.311
7	0.195	0.0825	0.00132	0.00867	0.137	0.208	0.265	0.309	0.311
10	0.202	0.0771	0	0.0152	0.148	0.215	0.267	0.309	0.311
Total	0.542	19.07	0	0.0121	0.144	0.213	0.266	0.310	7846.8

Distance measured in Miles. Part 8 and 9 are dropped because of uninformative.

Table 13: Distance by Parts

B.2 Group Summary

Group	# of Obs	Total	Percent
Group 1	547563	32457426	.0169
Group 2	228670	32457426	.007
Group 3	134891	32457426	.0042
Group 4	24127	32457426	.0007
Group 5	31522175	32457426	.9712

Table 14: Descriptive Summary (need to be updated)

Group	Mean	Std	Min	P1	P25	Median	P75	P99	Max
1	20.58	146.0	0	0.00313	0.0317	0.0955	0.430	685.6	7846.8
2	0.101	0.0865	0	0.00292	0.0277	0.0745	0.160	0.303	0.311
3	0.119	0.0897	0	0.00406	0.0405	0.0972	0.191	0.305	0.311
4	0.195	0.0825	0.00132	0.00867	0.137	0.208	0.265	0.309	0.311
5	0.202	0.0771	0	0.0152	0.148	0.215	0.267	0.309	0.311
Total	0.542	19.07	0	0.0121	0.144	0.213	0.266	0.310	7846.8

Table 15: Distance by Groups

Appendix C Result

C.1 Sensitivity Analysis

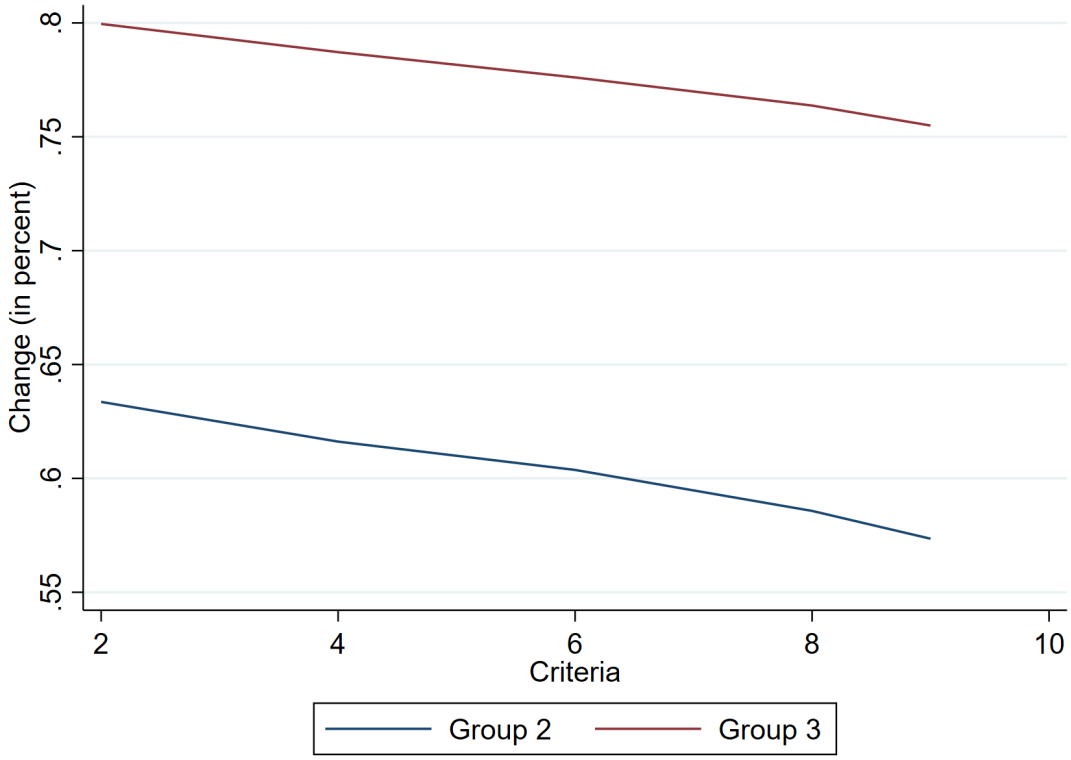


Figure 5: Sensitivity Analysis